Statistical Science
2019, Vol. 34, No. 2, 224–228
https://doi.org/10.1214/19-STS709
Main article: https://doi.org/10.1214/18-STS674

Comment: Empirical Bayes, Compound Decisions and Exchangeability

Eitan Greenshtein and Ya'acov Ritov

Abstract. We present some personal reflections on empirical Bayes/compound decision (EB/CD) theory following Efron (2019). In particular, we consider the role of exchangeability in the EB/CD theory and how it can be achieved when there are covariates. We also discuss the interpretation of EB/CD confidence interval, the theoretical efficiency of the CD procedure, and the impact of sparsity assumptions.

Key words and phrases: f-modeling, g-modeling, sparsity.

1. INTRODUCTION

We follow Efron (2019) in considering our perspective on the empirical Bayes (EB) research.

The empirical Bayes/compound decision (EB/CD, respectively) problem is the following:

$$\begin{aligned} &\theta_1,\dots,\theta_n \in \Theta \quad \text{unknown parameters} \\ &X_i \mid \boldsymbol{\theta} \sim F(\cdot \mid \theta_i), \quad i=1,\dots,n \text{ independently} \\ &\boldsymbol{L}(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}) = \sum_{i=1}^{n} L(\theta_i,\hat{\theta}_i), \end{aligned} \tag{1}$$

where bold symbol denotes the vector of the corresponding elements; thus $\boldsymbol{X} = (X_1,\dots,X_n)'$, $\boldsymbol{\theta} = (\theta_1,\dots,\theta_n)'$, etc. Saying it simply, we have n observations, each one on a different parameter. We want to estimate these parameters. The difference between the EB and CD *interpretations* is whether $\boldsymbol{\theta}$ is considered a vector of random variables with an unknown common distribution G, or just unknown parameters. There is very little technical difference between the two problems.

In all EB/CD problems, the statistic of interest is a vector of estimators of the individuals parameters $\theta_1,\dots,\theta_n$, where the loss function is additive in the loss of each of the n estimating problems. We emphasize this to distinguish the EB/CD setup from the general semiparametric inverse problems with mixing or with latent variables. These semiparametric problems, which have a rich history going back at least to the g-factor of intelligence of Spearman (1904), can be formulated as an inference about a parameter $\nu = \nu(G) \in R^k$ (possibly, $k=\infty$) after observing a sample from $X \sim \int F(\cdot \mid \theta)\,dG(\theta)$. In these problems, the final interest is in the (empirical) distribution of θ and not in the individual estimators $\hat{\theta}_i$, $i=1,\dots,n$. Thus, in our personal view, the species problem is not an EB/CD problem and will not discussed further in this comment.

In the plain vanilla EB problem, $\theta_1,\dots,\theta_n$ are i.i.d. random variables. In the corresponding CD problems, they are unknown parameters. A crucial element in both cases is that the order of $(\theta_1,X_1),\dots,(\theta_n,X_n)$ is arbitrary and noninformative. We consider such a situation as *exchangeable*. All estimators which respect this and neglect the order are *permutation equivariant*—satisfying

$$\hat{\boldsymbol{\theta}}(\pi \circ \boldsymbol{X}) = \pi \circ \hat{\boldsymbol{\theta}}(\boldsymbol{X}) \quad \forall \pi \in \Pi, \tag{2}$$

where Π is the set of all permutations of $1,\dots,n$ and for every $\pi \in \Pi$ and $a \in R^n$, $\pi \circ a = (a_{\pi(1)},\dots,a_{\pi(n)})'$.

Efron complains, "Considering the enormous gains potentially available from empirical Bayes methods, the effects on statistical practice have been somewhat underwhelming." The problem with the plain vanilla formula is that it is, well, plain. Some spices, nuts, or chocolate should be added to make it more useful.

Eitan Greenshtein, Ph.D., is with the Israel Central Bureau of Statistics, Kanfei Nesharim 66, 9546456 Jerusalem, Israel (e-mail: eitan.greenshtein@gmail.com). Ya'acov Ritov is Professor, Department of Statistics, University of Michigan, 1085 South University, Ann Arbor, Michigan 48109-1107, USA, and The Federmann Center for the Study of Rationality, The Hebrew University of Jerusalem, Edmund J. Safra Campus, 91904 Jerusalem, Israel (e-mail: yritov@umich.edu).

There are very few situations where we encounter n exactly similar problems with no extra information that makes them nonexchangeable. Usually the order is important, or there are other observed covariates which are different from an observation to an observation. The extensions we consider in the following are, naturally, those which we found as interesting research topics.

2. NONEXCHANGEABLE PROBLEMS AND THE PRESENCE OF COVARIATES

The main feature that the Robbins EB problems have is that the observations are exchangeable. The idea of CD is that it is possible to gain efficiency by learning from similar statistical problems. Similar is understood here as a formal statistical concept: the order of the observations $X_1, \ldots, X_n$ is not informative, and $\mathcal{L}(X_i \mid \theta_i = \theta)$ does not depend on i. This is a very restrictive structural assumption, which makes the approach rarely relevant to actual real problems where order is important and the distribution of X_i may depend on covariate Z_i in addition to θ_i.

It is not true that EB ideas are not standard tools of the trade. In fact, in some contexts they are the standard. It is a rare that one faces n exchangeable problems with n at least moderate. However, if the order is relevant, it is more natural to model $\theta_1, \ldots, \theta_n$ as a random process, for example, a Markov chain. The Kalman Filter and the general hidden Markov models (HMMs) are standard tools used to estimate $\theta_1, \ldots, \theta_n$ when the observations $Y_1, \ldots, Y_n$ are independent given $\theta_1, \ldots, \theta_n$ and $\mathcal{L}(Y_i \mid \theta_1, \ldots, \theta_n) = \mathcal{L}(Y_i \mid \theta_i)$. In the filtering problem, we estimate $\theta_1, \ldots, \theta_n$ with additive loss function $\sum L(\theta_i, \hat{\theta}_i)$, which is the essence of the EB problem minus the permutation invariance. More generally, a frequentist may consider $\theta_1, \ldots, \theta_n$ as fixed unknown constants, but he may want to minimize the risk when mostly $\theta_i \approx \theta_{i+1}$, $i = 1, \ldots, n-1$. For example, only a few change points are permitted, or any other shape constraint like monotonicity and boundedness are imposed on the vector $\boldsymbol{\theta}$.

Exchangeability is lost when there are covariates. For example, suppose that we know of two subsets of the measurements, one subset has measurements that were taken from males and the other has measurements that were taken from females. One might consider the measurements in each subset as exchangeable and apply empirical Bayes ideas separately on each subset.

Consider now a more general situation where there is a covariate Z_i independent of X_i given θ_i. One may split the data into K subsets based on their values of Z_i and apply EB/CD ideas separately on each subset (cf. Weinstein et al. (2018)). The finer is the split the lower is the Bayes risk $\sum_{k=1}^{K} \#\{Z_i : Z_i \in A_k\} R_{G(\cdot|Z\in A_k)}$, where R_G is the risk for estimating θ when the prior is G. The best would be to partition the sample to singletons. However, in the EB context, the Bayes procedure should be estimated given the data, and if we consider each sub-sample A_k separately, each sub-sample should be relatively large, so that the Bayes estimator can be learnt from the the observations in the stratum. Hence, unless n is extremely large, a fine partition yields a poor estimator.

In many cases, there is an alternative approach, which was discussed extensively in the literature. It is in line with Efron's equations (85) and (86) in the discussed paper. We extend this discussion somewhat and bring our perspective on the topic. Consider the classical case $X_i \sim N(\theta_i, \sigma^2)$. Let $\beta(Z_i)$ be a proxy for θ_i based on the covariate Z_i. Let $\tilde{X}_i = X_i - \beta(Z_i)$. Since it is assumed that $X_i \perp\!\!\!\perp Z_i \mid \theta_i$, then $\tilde{X}_i \mid \tilde{\theta}_i \sim N(\tilde{\theta}_i, \sigma^2)$, where $\tilde{\theta}_i = \theta_i - \beta(Z_i)$. This translates the non permutation equivariant original problem to the plain-vanilla EB/CD formulation (1) since the information in Z_i was exhausted by $\beta(Z_i)$. Let $\hat{\tilde{\theta}}_i$ be the CD estimator for this problem; then, one may use $\hat{\theta}_i = \hat{\tilde{\theta}}_i + \beta(Z_i)$ as the CD estimator in the original problem.

This approach was first suggested in the seminal paper of Fay and Herriot (1979), who considered the linear functions $\beta(Z_i) = \beta' Z_i$ and used the parametric-empircal-Bayes of Efron and Morris (1973) for solving the induced plain-vanilla EB problem. Cohen, Greenshtein and Ritov (2013) extended the idea to a nonparametric empirical Bayes estimation of θ_i, when the function β may be chosen from a class $\mathcal{B}$ of functions. The criterion for choosing an optimal β has to do with the corresponding distribution that is induced on the corresponding $\tilde{\theta}$. As demonstrated by Cohen et al., the heuristic and appealing approach of selecting $\beta_0 = \operatorname{argmin}_{\beta\in\mathcal{B}} \mathrm{E}(\theta - \beta(\mathrm{Z}))^2$ (e.g., the least squares estimator in the linear regression problem or the smoothing spline in the discussed paper) is not optimal. See their examples 3 and 4.

Greenshtein, Mansura and Ritov (2018) apply similar ideas to the estimation of $\theta_1, \ldots, \theta_n$ that are assumed tactically (and not necessarily correctly) to follow a state-space model. Assume that $(X_1, \ldots, X_n)$ are independent and normally distributed given $\theta_1, \ldots, \theta_n$. They considered two cases. In the first $Z_i = (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$, and, in the second, $Z_i = (X_1,$

$\ldots, X_{i-1})$, corresponding to in-line and off-line filtering. They took $\beta(Z_i)$ as the natural predictor of θ_i given Z_i (typically, the Kalman filter). Their estimator improves upon the standard Kalman filter estimator whenever the true model is not fully Gausssian.

3. NONPARAMETRIC f- AND g-MODELING

The nonparametric estimator considered by Efron is the Robbins estimator,

$$\hat{\theta}_i = (X_i + 1)\frac{N_n(X_i + 1)}{N_n(X_i)}, \tag{3}$$

where $N_n(x) = \sum_{i=1}^{n} \mathbf{1}(X_i = x)$. *Asymptotically*, it is actually quite good. Brown, Greenshtein and Ritov (2013) proved that its loss of efficiency relative to the Bayesian risk is only $O((\log(n)/\log\log(n))^2 n^{-1})$. However, its actual behavior is quite poor when n is small or moderate—the empirical ratio between two relatively small proportions is unstable. Brown et al. (2013) exemplify this fact empirically and suggest a few possible improvements to the basic estimator. The main steps include the smoothing and monotonizing the estimator. This follows since a proper estimator of f should satisfy some shape constraints. In particular, $\hat{\theta}(x)$ should be monotone increasing in x by the monotone likelihood property of the exponential families.

There is inherent inefficiency in the estimator (3). Its value at $X_i = x$ depends only on observations with values x and $x + 1$. Typically, this value x can be obtained from a wide range of values of θ, and all observations are informative about the underlined distribution of θ. This information is not used in evaluating $\hat{\theta}_i$, and the estimator may be very inefficient when the sample size is moderate. This motivates the other important nonparametric alternative—the g-modeling—the nonparametric estimation of g, for example, as suggested by Koenker and Mizera (2014). The decision process is automatically monotone, and the NPMLE is an off the shelf tool that can be generally implemented without too much difficulty. It is not clear to us that the NPMLE is an efficient estimator for our task (our anecdotal experience is that it is not, and an early stopping, as a way of smoothing the NPMLE, improves it considerably).

It is hard to compare theoretically the effect on estimation of these f and g nonparametric modeling. Part of the problem is that, as written above, theoretically speaking, the Robbins's estimator leaves very little for an asymptotic improvement.

4. EB CONFIDENCE INTERVALS

The EB statement is essentially non-Bayesian, certainly not in any subjective sense of this philosophy. Even if θ_i is random, it is random in the frequentist sense of the word and, hence, every statistician, whether he considers himself Bayesian or frequentist, may consider it as random. The "prior," the underlying empirical distribution of $\theta_1, \ldots, \theta_n$, has only the frequentist sense of probability. Ultra orthodox frequentists do accept Bayes theorem. They do not believe in subjective probabilities.

Efron raises the question whether there is a frequentist notion of EB confidence intervals. His answer is no, and EB intervals can have interpretation only as Bayesian credible sets. We think that as other aspects, EB confidence intervals can be understood mainly in the frequentist sense. Frequentists do have confidence intervals for random variables, for example, the confidence intervals for prediction in the standard linear model, or confidence bound in the filtering problem of state space models. Since the framework of the EB/CD is when $n \to \infty$, while we want to be informative for every θ_i, the familywise error rate may not be a relevant concept. The proper notion is that of *mean coverage*. Confidence sets C_i, $i = 1, \ldots, n$, with confidence $1-\alpha$ are such that $n^{-1}\sum P(C_i \ni \theta_i) > 1-\alpha$ whatever is the value of $\boldsymbol{\theta}$. This definition conforms to the logic that for any values of the parameter the confidence interval should include the true value, "on the average," in $(1-\alpha)100\%$ of the times it is calculated.

Nonparametric g-modeling is a convenient way to compute EB/CD confidence interval, the EB credible sets. The nonparametric MLE has typically a finite support (e.g., this is the case when an exponential family is assumed), hence the credible set for each θ_i should be something like the convex hull of the formal credible set with the NPMLE as the prior (if it is consistent).

It is possible to use the f-modeling as well. Here we present a naively simple example for a confidence bound. Efficient confidence sets would be more complex and are beyond the scope of this note. Suppose $X_i \sim N(\theta_i, \sigma^2)$, and suppose, for simplicity, that we want an $1-\alpha$ confidence bound. Following the Efron and Morris (1973) parametric EB approach, we suggest a confidence bound of the form $(\tau^2/(\sigma^2+\tau^2))X + c$, where τ^2 is the variance of $\theta_1, \theta_2, \ldots$. However, we want to find c such that the coverage would be true

even if the prior G is not normal. That is, we need

$$\begin{aligned}\alpha &= \int \Phi\left(\frac{1}{\sigma}\left(\frac{\sigma^2}{\tau^2}\theta - \frac{\sigma^2+\tau^2}{\tau^2}c\right)\right) dG(\theta)\\ &= \int \Phi\left(\frac{\sigma}{\tau^2}\left(\theta - \frac{\sigma^2+\tau^2}{\sigma^2}c\right)\right) dG(\theta)\\ &= F^*\left(-\frac{\sigma^2+\tau^2}{\sigma^2}c\right),\end{aligned}$$

where $F^* = G \star N(0, \tau^4/\sigma^2)$. But, assuming that the likelihood is more informative than the prior, $\tau^2 > \sigma^2$ and, hence, $F^* = F \star N(0, (\tau^4 - \sigma^4)/\sigma^2)$. Thus, F^* can be easily estimated by a kernel smoothing of the empirical distribution of X with a fixed $N(0, (\tau^4 - \sigma^4)/\sigma^2)$ kernel.

5. THE NATURE OF THE ORACLE

Orcales are introduced in statistics in order to prove the efficiency of a procedure. If we consider the EB setup (1) with $\theta_1, \ldots, \theta_n$ i.i.d. from an unknown G, the oracle is simple, that is, he is the Bayesian who knows G. Since the pairs (θ_i, X_i) are i.i.d., the Bayesian estimator has the form of $\hat{\theta}_i = \delta(X_i; G)$. We call a decision rule of the form $\delta_i(X_1, \ldots, X_n) = \delta(X_i)$ a separable rule. The EB procedure to be used is most likely not separable, but we can argue that in many important situations the Bayesian separable procedure can be well approximated based on the data.

The situation with CD is less simple—we do not have a theoretical G to consider. We may even assume that $\boldsymbol{\theta}$ was selected by an adversarial agent, who understands the statistician's procedure. A real oracle is needed to argue for the efficiency of the procedure. Three conditions are needed in order for an oracle argument to be valid: (1) The oracle should know everything the statistician knows; (2) he should be no more restricted than the statistician; and (3) he should be efficient given what he knows and what he is restricted to do.

To prove the efficiency of the standard procedures applied to the CD problem (1), Efron considers an oracle who knows G_n, the empirical distribution of $\boldsymbol{\theta}$, or, in other words, who knows $\theta_1, \ldots, \theta_n$ up to an unknown permutation. It is implicitly implied that the oracle would choose $\hat{\theta}_i$ as the Bayes procedure based on observing X_i and with prior G_n, $\hat{\theta}_i = \delta(X_i; G_n)$. In particular, if L is the standard quadratic loss function then

$$\hat{\theta}_i = \delta(X_i; G_n) = \frac{\int \theta f(X_i \mid \theta)\, dG_n(\theta)}{\int f(X_i \mid \theta)\, dG_n(\theta)}. \tag{4}$$

However, this is not that simple.

The parameter sets the oracle faces is the set of all $n!$ permutations of $\boldsymbol{\theta}$. The oracle who wants to ensure the minimax risk against an adversary, and naturally does not worry about computational complexities, will use the Bayes procedure with respect to the uniform probability distribution over the $n!$ permutations:

$$\hat{\boldsymbol{\theta}} = \frac{\sum_{\pi\in\Pi} \pi \circ \boldsymbol{\theta} \prod_{i=1}^n f(X_i \mid \theta_{\pi(i)})}{\sum_{\pi\in\Pi} \prod_{i=1}^n f(X_i \mid \theta_{\pi(i)})}. \tag{5}$$

This procedure is minimax since it is both Bayes and an equalizer.

Moreover, since the CD problem is invariant under permutation, it may seem natural for the oracle to use an equivariant procedure. The estimator suggested in (5) is the only possible Bayes (and hence admissible) equivariant procedure for the oracle, that is, $\mathrm{ASE}(\theta, \hat{\theta}) < \mathrm{ASE}(\theta, \Delta)$ for any other equivariant procedure Δ.

Generally speaking (4) and (5) are different. Suppose $n = 2$, the oracle is told that $\{\theta_1, \theta_2\} = \{-1, 1\}$, and $X_i \sim N(\theta_i, 1)$. In this case, the ASE of the permutation equivariant estimator of the oracle is 48% lower than that his separable procedure. For example, if $\boldsymbol{X} = (0, 4)'$. Then, a reasonable oracle will use $\hat{\boldsymbol{\theta}} \approx (-1, 1)'$ as prescribed by (5) and not $\hat{\boldsymbol{\theta}} \approx (0, 1)'$ as implied from (4). When $\theta_1, \ldots, \theta_n$ are i.i.d. $N(0, 9)$, the minimax risk is much lower than that of the separable estimator for n as large as 150, as can be seen in Figure 1. Of course, being humans, we couldn't compute the exact equivariant estimator but only approximated it, so the result of the simulations presented in Figure 1 are only an upper bound on the actual risk.

However, Greenshtein and Ritov (2009) strengthen the results of Hannan and Robbins (1955), and prove that under some mild conditions the two oracle's estimators are asymptotically equal up to $O(1/n)$ in terms of their ASE risk. Such accuracy is needed for dealing with sparse CD problems where the ASE of the (5) is $O(1/n)$. For example, if $X_i \sim N(\theta_i, 1)$, $\theta_i \in [a, b]$, $i = 1, \ldots, n$, the conditions are met and the estimator of (5) can be approximated by a separable procedure.

The asymptotic equivalence of (4) and (5) is necessary for establishing the asymptotic optimality of the EB/CD procedure. Hence, at least under some conditions, the oracle may use (4). But, the standard EB/CD arguments show that in many interesting models (4) can be well approximated by the statistician, and the EB/CD procedures achieves the minimal possible ASE.

This argument cannot be avoided. One could try to restrict arbitrary the oracle to separable decision procedures of the form $\tilde{\theta}_i = \tilde{\theta}_i(X_i; G_n)$. But the statistician

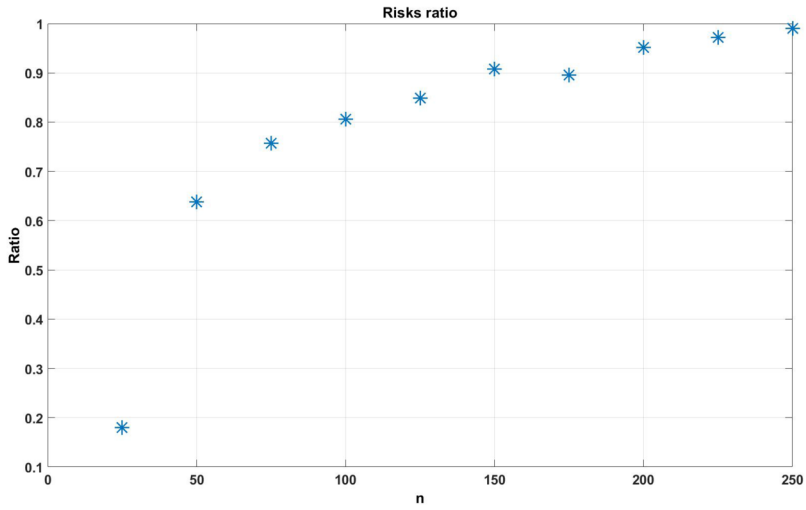

FIG. 1. *The ratio between the* (*approximated*) *risk of the equivariant oracle of* (5) *and the separable one of* (4) *as a function* n, *where* $\theta \sim N(0, 9)$ *and* $X \mid \theta \sim N(0, 1)$ *as a*.

does not know G_n, and the all nature of the EB/CD idea is that the statistician's $\hat{\theta}_i$ depends on all the observations, not only on X_i, $\hat{\theta}_i = \delta(X_i; X_1, \ldots, X_n)$, which is not a separable procedure. Thus, under this restriction, although the oracle knows more than the statistician, he has less freedom, and hence the two cannot be compared and this oracle is of no use.

That is all for these strange creatures, the oracles.

6. SPARSITY

Dealing with sparse vectors seems to be in the center of attention of the current research. When the vector $\boldsymbol{\theta}$ is sparse, CD procedures typically excel. When most of $\theta_1, \ldots, \theta_n$ are 0, NPMLE detects mostly noise and linear shrinkage shrinks all the way to 0. Many CD procedures automatically detect the range of the nonzero parameters and the conditional (empirical) distribution of $\{\theta_i : \theta_i \neq 0\}$.

However, this is a context where the right asymptotic formulation is that of triangular arrays. The Bayes formulation, when $\theta_i \sim G$ for G that does not depend on n cannot hold since the fraction of nonzero parameters converges to 0. It seems that the natural formulation is that of the compound decision problem. See Jiang and Zhang (2009), Greenshtein, Park and Ritov (2008), and Brown and Greenshtein (2009).

REFERENCES

BROWN, L. D. and GREENSHTEIN, E. (2009). Nonparametric empirical Bayes and compound decision approaches to estimation of a high-dimensional vector of normal means. *Ann. Statist.* **37** 1684–1704. MR2533468

BROWN, L. D., GREENSHTEIN, E. and RITOV, Y. (2013). The Poisson compound decision problem revisited. *J. Amer. Statist. Assoc.* **108** 741–749. MR3174656

COHEN, N., GREENSHTEIN, E. and RITOV, Y. (2013). Empirical Bayes in the presence of explanatory variables. *Statist. Sinica* **23** 333–357. MR3076170

EFRON, B. and MORRIS, C. (1973). Stein's estimation rule and its competitors—An empirical Bayes approach. *J. Amer. Statist. Assoc.* **68** 117–130. MR0388597

FAY, R. E. III and HERRIOT, R. A. (1979). Estimates of income for small places: An application of James–Stein procedures to census data. *J. Amer. Statist. Assoc.* **74** 269–277. MR0548019

GREENSHTEIN, E., MANSURA, A. and RITOV, Y. (2018). Nonparametric empirical Bayes improvement of common shrikage estimators. Submitted.

GREENSHTEIN, E., PARK, J. and RITOV, Y. (2008). Estimating the mean of high valued observations in high dimensions. *J. Stat. Theory Pract.* **2** 407–418. MR2528789

GREENSHTEIN, E. and RITOV, Y. (2009). Asymptotic efficiency of simple decisions for the compound decision problem. In *Optimality. Institute of Mathematical Statistics Lecture Notes—Monograph Series* **57** 266–275. IMS, Beachwood, OH. MR2681676

HANNAN, J. F. and ROBBINS, H. (1955). Asymptotic solutions of the compound decision problem for two completely specified distributions. *Ann. Math. Stat.* **26** 37–51. MR0067444

JIANG, W. and ZHANG, C.-H. (2009). General maximum likelihood empirical Bayes estimation of normal means. *Ann. Statist.* **37** 1647–1684.

KOENKER, R. and MIZERA, I. (2014). Convex optimization, shape constraints, compound decisions, and empirical Bayes rules. *J. Amer. Statist. Assoc.* **109** 674–685.

SPEARMAN, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology* **15** 72–101.

WEINSTEIN, A., MA, Z., BROWN, L. D. and ZHANG, C.-H. (2018). Group-linear empirical Bayes estimates for a heteroscedastic normal mean. *J. Amer. Statist. Assoc.* **113** 698–710. MR3832220